FIMM

REPRODUCIBILITY AND VALIDITY STUDIES

of
Diagnostic Procedures in Manual/Musculoskeletal Medicine

Protocol formats

THIRD EDITION

FIMM SCIENTIFIC COMMITTEE
Editor: J. Patijn, MD, PhD

Preface to the third Reproducibility and Validity Protocol

Based on an internal discussion within the Scientific Committee (SC) of the International Federation for Manual/Musculoskeletal Medicine (FIMM), a third protocol became necessary. It became clear that the second protocol showed shortcomings with respect to other aspects of kappa statistics, such as the weighted kappa, the value of the significance and confidence intervals of the kappa. New research results such as the effect of education on the kappa value and the proof of the 50% method to influence the prevalence in advance are incorporated in the present protocol. It became clear that there was a need to describe inappropriate statistics used in reproducibility studies in Manual/Musculoskeletal Medicine (M/M Medicine). More attention is paid to the different kind of characteristics of data collected in reproducibility studies, such as ordinal, nominal and interval or continuous data.

Based on experience with recent kappa studies, the format of the reproducibility protocol is adapted in several aspects of the different phases of a study.

This SC protocol in particular emphasises the kappa method for reproducibility studies of diagnostic procedures in M/M Medicine. For these kind of studies a "Cook Book" format is presented in a very practical way to make it available for both clinics with two or more physicians in M/M Medicine and Educational Committees of National Societies to perform these kind of studies.
For university departments more in-depth information about statistics in every kind of reliability studies is provided in this SC protocol.

The Scientific Committee of the FIMM is aware that developing this kind of protocols is a continuous process.
By publishing the third protocol on the website of the FIMM, the Scientific Committee hopes that those scientists who use this protocol will send their comments to the Chairman of the Scientific Committee. In this way, we hope to improve the present protocol.

The SC asks those scientists who receive this protocol to distribute this protocol to their fellow scientists. In this way, the protocol

becomes accessible for all practitioners in the field of M/M Medicine.
This protocol is the end product of the energy of all members of the SC.

Dr. Jacob Patijn, MD, PhD, Neurologist,
Physician for Manual/Musculoskeletal Medicine
Chairman of the Scientific Committee of the FIMM
Responsible member for the Reliability Group of this Committee

SCIENTIFIC COMMITTEE FIMM

Chairman, Dr. Jacob Patijn, Eindhoven, The Netherlands

Members:
Dr. Jan van Beek, MD, The Hague, The Netherlands
Dr. Stefan Blomberg, MD, PhD, Stockholm, Sweden
Professor Boyd Buser, DO, Biddeford, United States
Dr. Richard Ellis, MD, PhD, Salisbury, United Kingdom
Dr. Jean Yves Maigne, MD, PhD, Paris, France
Dr. Ron Palmer, MD, Herston, Australia
Dr. Lars Remvig, MD, Holte, Denmark
Dr. Jan Vacek, Prague, MD, Czech Republic
Professor Robert Ward, DO, Michigan, United States
Professor Lothar Beyer, MD, PhD, Jena, Germany
Professor Olavi Airaksinen, MD, PhD, Kuopio, Finland

Advisor:
Professor Dr. Bart Koes, PhD, Epidemiologist, Erasmus University Rotterdam

Address for reprints and comments:
Dr. Jacob Patijn, MD, PhD, Neurologist
University Hospital Maastricht, Pain Management and Research Centre, Dept. Anaesthesiology, Maastricht, The Netherlands, Fax: 31 43 3875457, E-mail jpat@sane.azm.-nl, jacobpatijn@hccnet.nl

FIMM

International Federation for Manual/Musculoskeletal Medicine

I. INTRODUCTION CHAIRMAN SCIENTIFIC COMMITTEE

I.1 Background

This is the fifth of the protocols published by the Scientific Committee (SC) of FIMM.

Its concerns a standardised format for validity, sensitivity and specificity studies. Besides, it provides the scientist as well as the daily practitioners in our field in more or less cook book form with a format for reproducibility studies in Manual/Musculoskeletal Medicine.

In the future, continuously improved scientific committee protocols will be developed.

The reason of the SC to develop these kind of protocols has been extensively discussed in previous reports of the SC for the General Assembly and has been published in FIMM NEWS. In different countries, the previously published protocols gradually have led to reproducibility studies in M/M Medicine.

The primary reason to develop this kind of protocols by the SC is still actual. Therefore, as we did in previous protocols, a short background is provided of these protocols and a brief overview of the past SC activities is included.

The Scientific Committee of FIMM (SC) formulated the problem with respect to diagnostic procedures in Manual/Musculoskeletal Medicine (M/M Medicine), and it is summarised in the statement:

There are too many different schools in Manual/ Musculoskeletal Medicine in many different countries of the world, with too many different diagnostic procedures and too many different therapeutic approaches.

The consequences of this statement are five-fold:

I.1 Most schools within M/M Medicine have not validated yet their own characteristic diagnostic procedures in the different regions of the locomotion system. Therefore reproducibility, validity, sensitivity and specificity of these diagnostic procedures are still lacking.

I.2 All the different schools within M/M Medicine still coexist. Because of lack of good reproducibility, validity, sensitivity and specificity studies, mutual comparison of diagnostic procedures is impossible. Scientific information exchange and fundamental discussions between these different schools, based on solid scientific methods, are hardly possible in the present situation.

I.3 Absence of validated diagnostic procedures in M/M Medicine leads to heterogeneously defined populations in efficacy trials. Therefore, comparison of efficacy trials, with the same therapeutic approach (for instance manipulation), is impossible.

I.4 If the present situation is allowed to continue, it will lead to a slowing down of the badly needed process of professionalisation of M/M Medicine.

I.5 Non-validated diagnostic procedures of different schools, ill-defined therapeutic approaches and low quality study designs are the main causes for the weak evidence of a proven therapeutic effect of M/M Medicine.

It is still the opinion of the SC that the committee should create conditions for exchange of scientific information between the various schools in M/M Medicine. This information exchange must be based on results of solid scientific work. By comparing the results of good reproducibility, validity, sensitivity and specificity studies, performed by different schools, a fundamental discussion will arise. The main aim of this discussion is not to conclude which school has the best diagnostic procedure in a particular area of the locomotion system, but to define a set of validated diagnostic procedures which can be adopted by the different schools and become transferable to regular medicine.

The SC wants to provide the National Societies of FIMM with standardised scientific protocols for future studies.

The SC thought that the best forum for creating a discussion platform would be to organise every other year a SC Conference in cooperation with a particular National Society. The SC Conference was organised in Odense, Denmark, 2003, in cooperation with the Danish Society for Manual Medicine. Many researchers presented their preliminary results, proposals for protocol formats and therapeutic algorithms. In a fruitful discussion between audience and presenters many ideas were exchanged based on solid scientific work, without interference of “school politics”.

As Chairman of the SC, I want to emphasise that good reproducibility, validity, sensitivity and specificity studies still have the first priority. This kind of studies is easy and cheap to perform, and they form the best base for mutual discussion between schools in M/M Medicine.

Co-operation and active involvement of the National Societies of FIMM is indispensable and crucial for the future work of the SC.

In providing this third protocol to the National Societies of FIMM, the SC hopes to add a substantial contribution to the professionalisation of M/M Medicine.

Dr. Jacob Patijn, MD, PhD, Neurologist

II. REPRODUCIBILITY AND VALIDITY

Nomenclature

One of the major problems in medicine and in research is the fact that different names are used for the same definition Therefore we thought it important first to provide the reader of this protocol with an overview of the definitions used in this protocol. In clarifying the definitions in advance we hope to make reading easier.

II.1 **Reliability** can be divided in **Precision** and **Accuracy.**

II.1.1 **Precision,** also called **Reproducibility**

In the case of reproducibility of an observation made by one observer on two separate occasions, we call it the *intra-observer variability* or the *intra-observer agreement.*

In the case of reproducibility of an observation by two observers on one occasion, we call it the *inter-observer variability* or the *inter-observer agreement.*

In this protocol, we use the terms **reproducibility, intra-observer agreement** and **inter-observer agreement.**

Reproducibility of diagnostic procedures in M/M Medicine evaluates whether two observers find the same result of a diagnostic procedure in the same patient population, or whether a single observer finds the same result of a diagnostic procedure in the same patient population on two separate moments in time.

II.1.2 **Accuracy,** also called **Validity**

In this protocol, we use the term **validity.**

Validity measures the extent to which the diagnostic test actually does what it is supposed to do. More precisely, validity is determined by measuring how well a test performs against the gold or criterion standard.
When a diagnostic test has to be evaluated with respect to what it is supposed to do (validity), a gold standard as reference is needed. This is a major problem not only in M/M Medicine but in the whole medical profession. Sometimes, radiological findings, post-mortem findings or findings during an operation can act as gold standard. In the case of subjective quantification of range of motion, the gold standard can be the result of a quantitative method performed in a normal population. Frequently, it is only possible to define a gold standard by consensus of experts in a particular field of medicine.

Gold standards are needed for estimation of the sensitivity and specificity of a test (see V.1).

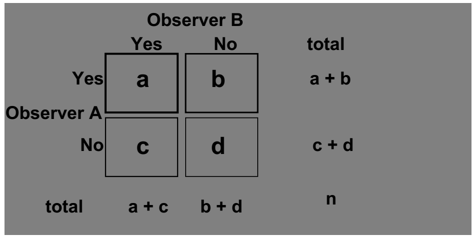
II.2 Index Condition and its Prevalence

II.2.1 The **index condition** is synonymous with the diagnosis of a patient. This diagnosis must be based on reproducible diagnostic procedures with a proven validity.
In case of reproducibility studies of diagnostic procedures, a positive judged test by observers is called the index condition.

II.2.2 The **prevalence of the index condition** is the frequency of the index condition in a particular population at a particular moment. In reproducibility studies of tests, the prevalence of the index condition is only related to the study population.
It is essential to realise that the prevalence of an index condition can vary in different institutes, countries and can change in time.

In the reproducibility section of this protocol, we will use the terms **index condition** and **prevalence of the index condition in relation to positive found test procedures.**

In the 2 x 2 contingency table hereunder, a theoretical example of the results of a reproducibility study of two observers A and B is shown.



		Observer B		
		Yes	No	total
Observer A	Yes	a	b	a + b
	No	c	d	c + d
total		a + c	b + d	n

Figure 1. 2 x 2 contingency table

The squares with a and b represent the number of patients with positive tests as judged positive by observer A. The squares with a and c represent the number of patients with positive tests as judged by observer B. The squares a, b and c represents the number of patients with positive tests as judged by either one or both observers among the total patients n.
The prevalence is calculated by the formula for the prevalence (P):

$$P= \frac{[a + (b + c)/2]}{n} \qquad \textbf{(formula 1)}$$

II.3 Overall Agreement

The overall agreement reflects the percentage of the patients in which both observers A and B agree about the judgement of the

test. Based on figure 1, both observers agree in a and d (respectively positive and negative). In the squares with b and c, the observers disagree.
Overall agreement P_o is calculated by the formula:

$$P_o = \frac{[a + d]}{n} \qquad \textbf{(formula 2)}$$

II.4 Sensitivity and Specificity

II.4.1 The **sensitivity** of a test is defined as the proportion of the cases that have the index condition that the test correctly detects.

II.4.2 The **specificity** of a test is defined as the proportion of the cases that do not have the index condition that the test correctly detects.

In this protocol the so-called "Nosographic Sensitivity and Specificity" is identical with the terms "Sensitivity and Specificity".

II.4.3 To translate the statistics of sensitivity and specificity figures into daily practice, the physician has to know whether a positive test in the individual patient is truly positive as opposed to false-positive. This is expressed respectively as the so-called "*positive predictive value* of a test" and "*negative predictive value* of a test".

In contrast to the "Nosographic Sensitivity and Specificity", the positive predictive value of a test and negative predictive value of a test are also called the "Diagnostic Sensitivity and Specificity".

In this protocol the so-called "Diagnostic Sensitivity and Specificity" is identical with the terms "**positive and negative predictive value** of a test".

II.5 Kappa Value: Interpretation

In this protocol, kappa statistics will be the method of choice for reproducibility studies (see below).
Kappa value is a statistical measurement for the intra-observer and inter-observer agreement corrected for chance. The kappa value can be either negative or positive and ranges between –1 and +1.
Several schemes are available (see Haas 5,6,12) to draw the line on good agreement. The most widely used scheme is that of Landis and Koch. They stated that kappa values above 0.60 represent good to excellent agreement beyond chance between two raters. In contrast, kappa values of 0.40 or less represent poor agreement beyond chance. Kappa values between 0.40 and 0.60 reflect a fair to good agreement beyond chance.

Bogduk uses a kappa value of 0.4 as cut off level of good agreement.
In this protocol we use a conservative kappa value cut off level 0.6, reflecting a good to excellent agreement.

III. STARTING POINTS IN REPRODUCIBILITY PROTOCOL OF DIAGNOSTICS IN M/M MEDICINE

To perform reproducibility studies for diagnostics in M/M Medicine, several points are important to consider to start with.

III.1 Character of the Diagnostic Procedure and Statistical Methods

Before starting a reproducibility study in M/M Medicine, it is important to be clear about what kind of diagnostic procedure we are dealing with and what kind of statistics are appropriate.
In general we have two kinds of diagnostic procedures: a. Qualitative Diagnostic Procedures, b. Quantitative Diagnostics Procedures.

III.1.1 Qualitative and Semi-Quantitative Diagnostic Procedures (Nominal and Ordinal Data)

Qualitative diagnostic procedures in M/M Medicine are characterised by subjective outcomes of observer and/or patient. These kinds of procedures can have both a nominal or an ordinal character. Typical examples of this kind of procedure in M/M Medicine are end feeling and pain provocation under different conditions (provoked by observer, provoked by movements of the patient). In case of existence or absence of a finding (Yes/No), for example pain provoking tests, we are dealing with nominal data and kappa statistics indicated. If different categories (with a natural order) of a test procedure can be distinguished, for example: no end feel, soft end feel and hard end feel and very hard end feel, we are dealing with ordinal data, and weighted kappa statistics are indicated. Also, semi-quantitative diagnostic procedures in M/M Medicine are in essence a qualitative diagnostic procedure with a dichotomous character. Typical examples of these kinds of semi-quantitative diagnostic procedures in M/M Medicine are measurement of left/right difference in subjective range of motion of the examiner (difference in range of motion, Yes or No or restricted motion, Yes or No).

III.1.2 Quantitative Diagnostic Procedures

In quantitative diagnostic procedures, mostly measured with a certain kind of device, findings are quantified in degrees, millimetres, kg etc. and are mentioned interval or continuous data.
For these kind of quantitative procedures normative values are needed. First a study of the procedure in normal subjects is

needed in which the reproducibility of the procedure has to be estimated in the same population on two different occasions. In this test/retest study, the systematic measurement failure can be estimated based on the distribution of the data values. Besides, factors such as age and gender, which can influence the data, have to be studied. Quantitative diagnostic procedures can serve as gold standard for semi-quantitative diagnostic procedures.

In reproducibility studies of any kind, the nature of the collected data (nominal, ordinal, interval or continuous) is decisive for the applied statistical method.

III.1.3 Inappropriate Statistics in Qualitative Data Reproducibility Studies

Frequently, inappropriate statistics are applied to measure the reproducibility. The main flaw is that agreement is often confused with trend or association, which is the assessment of the predictability of one variable from another. Hereunder the flaws of several statistical methods in reproducibility studies are listed.

III.1.3.1 Percent Agreement

Reproducibility studies, just mentioning the *percent agreement,* give no real information about the reproducibility. *Percent agreement* is the ratio of the number of subjects in which the observers agree to the total number of observations. The main problem is that the *percent agreement* does not take into account the agreement that is expected to occur by chance alone.

III.1.3.2 Correlation Coefficients

In many reproducibility studies correlation and association measures are used to evaluate the reproducibility of clinical data. The problem is that some do not have the ability to distinguish trend toward agreement from disagreement (Chi-Square [^{2}] and Phi) or do not account for systematic observer bias (Pearson's product moment correlation, Rank order correlation).

III.1.4 Appropriate Statistics in Qualitative and Semi-Quantitative Data Reproducibility Studies

III.1.4.1 *Normal Kappa* is the statistics of choice for evaluating reproducibility between two observers for nominal (dichotomous) data.

III.1.4.2 In case of many observers (>2) the *overall kappa* can be used to generalise the results to broader populations of observers. For example, evaluating the existence of segmental dysfunctions in a particular area as indication for therapy, the overall kappa would

give an estimate of the overall reproducibility to detect segmental dysfunctions by observers in that particular area. For details see textbooks or ask your statistical expert.

III.1.4.3 In M/M Medicine the judgement of a diagnostic procedure can be subdivided into different grades, such as end feel (normal, elastic, hard). These ordinal data must have a natural order. In reproducibility studies with ordinal data the statistics of *weighted kappa* is indicated. For details see textbooks or ask your statistical expert.

III.1.4.4 Significance of the found kappa value, together with confidence intervals, can be calculated, in case of kappa values between 0.40 and 0.60. It provides you with the information whether the found kappa value differs from chance. In case of kappa values over 0.60, this procedure is not necessary.

The ins and outs of normal kappa statistics is elaborated more in detail below (see III).

III.1.5 Appropriate Statistics in Quantitative Data Reproducibility Studies

To evaluate the reproducibility of measurements with quantitative data (interval or continuous data) in repeated measures, the paired t-test is indicated.
One-way analysis of variance intraclass coefficient (ANOVA ICC) is the statistical method of choice for the reproducibility of observers for interval data (cm, mm, etc). The calculated factor R in this statistical procedure is 1 if there are identical ratings, less than 0 in absence of reproducibility. A limitation of the ICC is that it provides no information about the magnitude of disagreement between observers.

In reproducibility studies, the choice of statistics should depend not only on the character of the collected data (nominal, ordinal, interval), but also on the related type of clinical decision concluded from the findings of the study.
For instance, if one needs the findings of the study to decide whether or not a heel lift is indicated to correct leg length inequality, ANOVA ICC statistics for interval data are indicated. In contrast, if leg length differences are measured to adjust pelvic adjustment, the data characteristics are right, left and equal and therefore kappa statistics are indicated for the nominal data. The same is true for semi-quantitative data such as the side of restricted range of motion Yes or No.

III.2 Aim of the Diagnostic Procedure

In studying the reproducibility of diagnostic procedures in M/M Medicine, one has to be clear about the aim of the test(s). It is essential to realise the difference between a diagnosis, a syndrome and a diagnostic test used in daily practice. In a genuine diagnosis, the aetiology and prognosis is known. In syndromes, a combination of signs and symptoms that appear together in a high frequency in a certain population, the aetiology is unknown.

In both diagnosis and syndromes, diagnostic tests are needed. A diagnostic test is a procedure, performed by a clinician, to objectify in a qualitative way a clinical finding which is frequently not mandatory with a genuine diagnosis. For example the combination of sensory deficit, motor deficit and a positive Lasègue can be characteristic for a radicular syndrome. The aetiology can be as well an intervertebral disc protrusion as a tumour in the intervertebral foramen, both with root compression.

In M/M Medicine educational systems, many tests are taught to the student as a procedure, for instance passive cervical rotation. The student just learns how to perform the whole procedure of passive cervical rotation (setting of the hand, applied force etc.). The explanation for such a restriction can have many reasons and therefore gives no information about a diagnosis.

Therefore, the first priority is to make the procedures with their judgements of all kinds of tests in M/M Medicine reproducible. In second instance find gold standard to validate these procedures. For example: the finding of a restricted cervical rotation (Yes or No) must be validated by a quantitative method with a specially designed device that measures the rotation in degrees in different age and gender groups.

Subsequently, reliable tests (reproducible and validated) can be use to define syndromes in M/M Medicine.

Finally, and often very difficult, gold standards have to be found for validation.

III.2.1 Evaluating a single diagnostic test only gives information about the reproducibility of the whole *test procedure*.

In the vast majority of single diagnostic tests, no information is obtained about a specific diagnosis based on that single diagnostic test and consequently no indication for a specific therapy is provided.

Therefore, a single diagnostic test seldom differentiates between normal subjects and patients. In general, in the absence of a gold standard, sensitivity and specificity studies are useless if they are based on a single reproducible diagnostic test.

III.2.2 Evaluating a combination of test procedures gives information no more than the reproducibility of these combinations of the tests. Without a gold standard, reproducible combinations of tests have no diagnostic value and can be seen in specific diagnosis and non-specific pain syndromes, but also in normal subjects. The disadvantage of testing many tests at the same time in reproducibility studies is the potential mutual dependency of the tests.

III.2.3 Reproduction of a test in time (perform the same diagnostics in the same patient after a time interval) can be used to estimate the sensitivity and specificity of a test. Such tests, when combined with other clinical data, can increase the ability to differentiate between patients and normal subjects. However, in the vast majority of cases, no information is obtained regarding a specific diagnosis based on this combination. In general, it is only in the presence of a gold standard that it will be useful to perform sensitivity and specificity studies, based on a combination of valid test procedures.

III.3 Number of Tests to be Evaluated

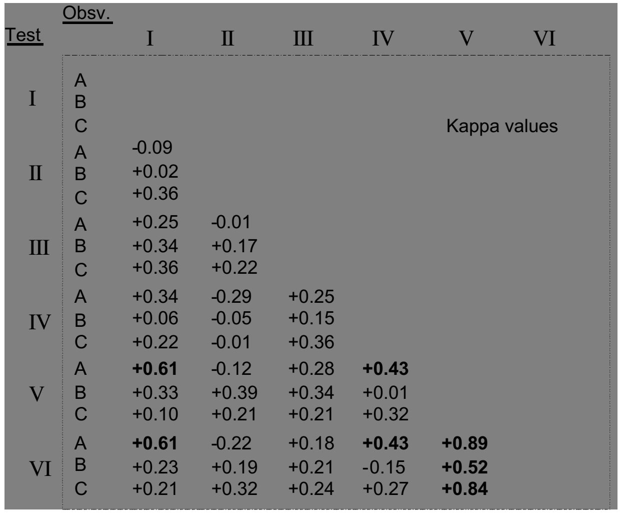
Reproducibility studies in non-specific, for instance in low back pain, sometimes show evaluation of reproducibility of a large number of tests at the same time. In this kind of studies, many of the tests show low kappa values and therefore are judged of no clinical importance by the authors. Since prevalence and overall agreement figures are frequently lacking, a definite conclusion about the reproducibility of the tests with low kappa values cannot be drawn. Heterogeneous study populations consist most probably out of different subgroups each with different prevalences of the tests to be evaluated. This can result in the risk that some positive tests have a low prevalence in the study, because of a small size of that particular unknown subgroup.

The tests to be evaluated must have a relation with the characteristics of the study population. For example, evaluating the reproducibility of several radicular provocation tests in LPB patients without any signs of sciatica has no sense, because it is to be expected that positive radicular tests are rare in such a population.

In case of a population with sciatica, evaluating the reproducibility of several radicular provocation tests at the same time, one can decide on a minimal number of positive tests which is needed to make the diagnosis of a lumbar radicular syndrome. The disadvantage of evaluating a combination of tests for a particular diag-nosis (for example radicular syndrome, SI-dysfunction) is that there is a chance for mutual dependency.

For example, many SI-tests in M/M Medicine are supposed to test a SI-dysfunction or hypomobility of the SI-joint. This mutual dependency was shown in a reproducibility study of six SI-tests at the same time (Deursen van, Patijn). In this study, three observers (A, B, C) were supposed to use six different SI-tests (I to VI) for the final SI-diagnosis (see figure 2 and 3).

To evaluate the mutual dependency of the tests, for each observer, the kappa values were calculated of the fifteen possible combinations of pairs of their six SI-tests.

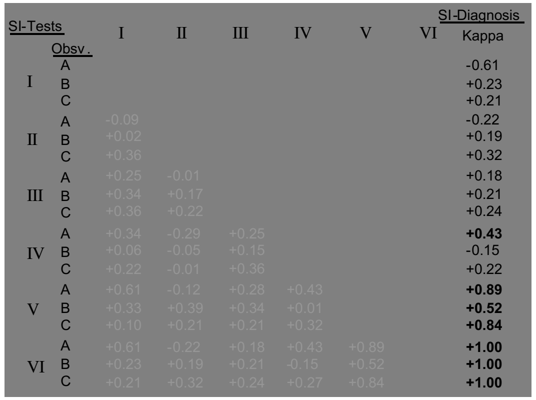


Test	Obsv.	I	II	III	IV	V	VI
I	A						
	B						
	C					Kappa values	
II	A	-0.09					
	B	+0.02					
	C	+0.36					
III	A	+0.25	-0.01				
	B	+0.34	+0.17				
	C	+0.36	+0.22				
IV	A	+0.34	-0.29	+0.25			
	B	+0.06	-0.05	+0.15			
	C	+0.22	-0.01	+0.36			
V	A	**+0.61**	-0.12	+0.28	**+0.43**		
	B	+0.33	+0.39	+0.34	+0.01		
	C	+0.10	+0.21	+0.21	+0.32		
VI	A	**+0.61**	-0.22	+0.18	**+0.43**	**+0.89**	
	B	+0.23	+0.19	+0.21	-0.15	**+0.52**	
	C	+0.21	+0.32	+0.24	+0.27	**+0.84**	

Figure 2. Mutual dependency of six SI-tests (I till VI) in three ob-servers A, B and C. The bold kappa values >0.40 reflect a mutual dependency.

Using a kappa of 0.40 as lowest level, figure 2 shows in different pairs of tests in all three observers a kappa value larger than 0.40. In particular between test V and VI, all observers showed high kappa values (+0.89, +0.52 and +0.84), reflecting a mutual dependency between test V and VI.
This means that all three observers unconsciously judged SI-test VI positive after they had judged SI-test V as positive. In this study, SI-tests II, III versus I, IV and VI show mutual independency (second and third column).

This aspect of mutual dependency is also very important in reproducibility studies when selecting tests for the same clinical feature/diagnosis. In kappa studies, besides evaluating the reproducibility of the tests themselves, the interobserver agreement of the final diagnosis, based on these tests, can be evaluated.
From the same study, as mentioned above, it became clear that with too many tests, observers use only a few tests for their final SI-diagnosis. By calculating the mutual kappa value of the single tests (I to VI) and the final diagnosis in all three observers A, B, and C this phenomenon is illustrated (see Figure 3).



								SI-Diagnosis
SI-Tests	Obsv.	I	II	III	IV	V	VI	Kappa
I	A							-0.61
	B							+0.23
	C							+0.21
II	A	-0.09						-0.22
	B	+0.02						+0.19
	C	+0.36						+0.32
III	A	+0.25	-0.01					+0.18
	B	+0.34	+0.17					+0.21
	C	+0.36	+0.22					+0.24
IV	A	+0.34	-0.29	+0.25				**+0.43**
	B	+0.06	-0.05	+0.15				-0.15
	C	+0.22	-0.01	+0.36				+0.22
V	A	+0.61	-0.12	+0.28	+0.43			**+0.89**
	B	+0.33	+0.39	+0.34	+0.01			**+0.52**
	C	+0.10	+0.21	+0.21	+0.32			**+0.84**
VI	A	+0.61	-0.22	+0.18	+0.43	+0.89		**+1.00**
	B	+0.23	+0.19	+0.21	-0.15	+0.52		**+1.00**
	C	+0.21	+0.32	+0.24	+0.27	+0.84		**+1.00**

Figure 3. Mutual dependency of six SI-tests (I till VI) with the final SI-diagnosis in three observers A, B and C. The bold kappa values > 0.40 reflect a mutual dependency.

Note that in the far right column “SI-Diagnosis”, all three observers only use SI-test V and VI for their final judgement of the SI-diagnosis. In all three observers A, B and C SI-tests I to IV contributed not at all to their final SI-diagnosis.
In general it is advisable to evaluate a maximum of three tests for the same clinical feature. It is advisable to choose tests each with a completely different procedure and not related to a single joint.

III.4 Number of Observers

There is no real statistical reason for performing a reproducibility study with more than two observers. In some studies, more observers are involved to evaluate the effect of the observers’ experience on the interobserver agreement. The problem with experienced observers is that they probably have developed a personal performance and interpretation of the test. Most of these studies lack a proper training period for standardisation of the performance of the test procedure and its interpretation. The results of these kinds of studies inform us more about the skills and/or the quality of the educational systems of the observers, rather than about the reproducibility of the evaluated tests. The same is true for reproducibility studies which estimate kappa values of tests done in the so-called “in-vivo condition”, in which no standardisation of the test procedures was carried out (to mimic the daily practice of a test).
The only case in which more observers can participate in kappa studies is to evaluate the effect of regular training on the kappa value. The same observers are repeatedly trained in a diagnostic

procedure and after each training period a new kappa is estimated to see whether a rise in kappa value in observers is seen.
In principle reproducibility studies, using the proposed format as discussed below, provide us with the potential reproducibility of a test procedure. If the reproducibility of a test procedure is established, a second study can be performed to evaluate the effect of observers' characteristics on the reproducibility.

A second flaw of using too many observers in a reproducibility study is the possibility of a therapeutic effect of the test procedure. If in a single patient, a passively performed procedure (passive cervical rotation) is performed too many times by different observers in a row, a therapeutic effect of the procedure may influence the range of motion and therefore the results of the last observer.

In general, using the proposed format in this protocol, two observers are sufficient to estimate the potential reproducibility of a test.

III.5 Hypothesis of a Test

It is very important for a reproducibility study of a test to discuss and analyse what the test is supposed to test. For range of motion there is no problem. For mobility, for instance hypomobility of the SI-joint, there is a problem. In many reproducibility studies of the SI-joint, the hypothesis for the various tests was that they were supposed to test the mobility of the SI-joint. Although SI-mobility is proven, based on cadaver studies, it is impossible, even for the most experienced observer, to test manually the mobility of the SI-joint. This incorrect belief is probably the reason for the low kappa values of SI-tests in the literature. Looking critically at the substantially different procedures of the large number of SI-tests, we have to question whether all these procedures can test the hypomobility of the SI-joint. In reproducibility studies, the observer has to forget the hypothesis of the tests taught by his teachers and has to concentrate on all the different aspects and details of the test procedure as such. For instance, according to the literature, the Patrick test for the SI-joint is supposed to test the mobility of a SI-joint. Looking critically at the test procedure, the observers can decide that the Patrick test, measuring end feeling and motion restriction, only evaluates increased muscle tension of a certain group of muscles related to the hip joint.
The effect of the hypothesis for the reproducibility on SI-tests was illustrated in two studies (Patijn 2000). The first study, which assessed six SI-tests supposed to evaluate SI-mobility, resulted in very low kappa values. In the second study, three tests supposed to test muscle hypertonia and its consequent motion restriction, in different muscle groups around the lumbosacral-hip region, resulted in a kappa value of 0.7.
Whatever tests one selects for a reproducibility study, one has to investigate step by step the whole test procedure and agree about what the test really tests.

Based on this agreement, the observers can define a more plausible hypothesis for the test, which can completely contradict the hypothesis stated in the literature.
Full agreement of the observers about a more plausible hypothesis of a test can lead to better results in reproducibility studies. In reproducibility studies these aspects are essential in the training period of the study format (see figure 10, page 27).

III.6 Blinding Procedures

In every reproducibility study, blinding procedures are essential not only for the patient/observer condition but also for both observers and must be well defined. Be sure that during the study there is no communication between observers, use separate forms for the observers to record their findings. If necessary, be sure there is no communication between observer and the patients.

III.7 Test Procedure and Test Judgement

As already argued under item 5, the observers have to standardise the whole test performance and the way they judge the result of a test. In the protocol format discussed below (see figure 10, page 27), the training period is essential for standardisation in a reproducibility study. The consensus about the definition of the test procedure and its assessment must be discussed in the final publication. To prevent observers' "personal interpretation" during the study, we also advise that the standardised procedures and test assessments are printed on the forms used in the study.

III.8 Selection and Number of Subjects

In reproducibility studies, the primary source population out of which the subjects are selected must be defined and mentioned in the final publication. Selection procedures must be very clear.
In general, for simple reproducibility studies 40 subjects are sufficient. This number of subjects makes this kind of reproducibility study easy and cheap to perform and not restricted to large institutes.

III.9 Statistics in Reproducibility Studies: the Kappa Value

In reproducibility studies with two observers evaluating dichotomous tests (Yes/No), estimation of the kappa values is the method of choice (see below).

III.9.1 Kappa Dependency on Prevalence

In the literature many reproducibility studies judge diagnostic tests with kappa values below 0.6 as clinically irrelevant. However, in the

vast majority of reproducibility studies no information is pre-sented about the corresponding prevalence and overall agreement of the index condition. This is essential, because the kappa value is dependent on the prevalence and the overall agreement.
Published reproducibility studies which present evaluations of tests with low kappa values, as clinically worthless or of minor importance, without mentioning any figures about prevalence and overall agreement, are misleading.

Low kappa values can reflect high as well as low prevalences!!!

Figure 4 shows the dependency of the kappa value on the prevalence.

Note that in case of very low (a) and very high prevalences (b) the kappa value becomes very low.

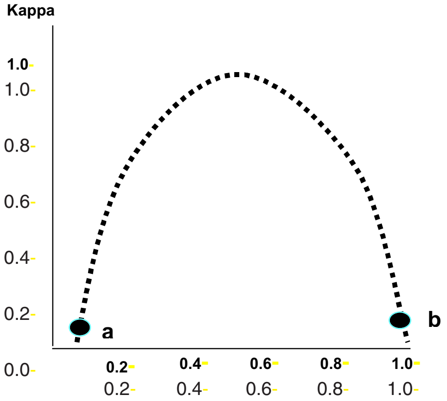


Figure 4. Relation between kappa values and prevalences

III.9.2 Kappa Dependency on Overall Agreement (P_o)

In figure 5 it is illustrated that with a high overall agreement (0.98 in the figure) the maximal kappa value is 1.0 and the minimal kappa value is nearly 0.

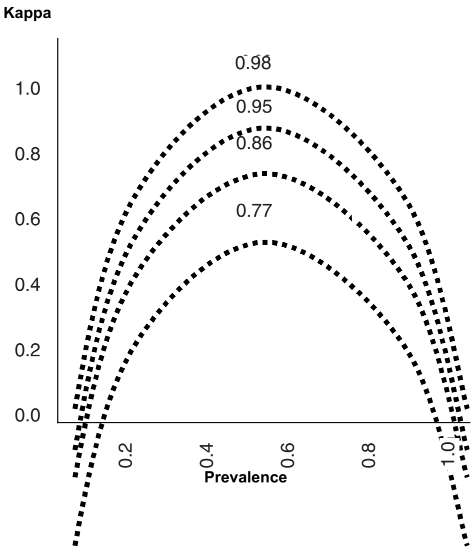


Figure 5. Relation between the different kappa/prevalence curves and the different Overall Agreements ranging from 0.77 to 0.98

The level of kappa values is dependent on the overall agreement P_o of the two observers. The lower the overall agreement in a reproducibility study, the lower the maximal and minimal kappa values become. In figure 5 this relation is shown. Note that in the prevalence/kappa curves with a low overall agreement P_o (0.86 and 0.77), the minimal kappa values become negative.

The dependence of the kappa value both on the prevalence P and on the overall agreement P_o illustrates the fact that a kappa value can only be interpreted in a proper fashion when both prevalence and overall agreements are mentioned in the final publication.

III.9.3 Optimising Procedures for Reproducibility Studies: Influencing the Overall Agreement and Prevalence in Advance to a Level of 50%

When performing a reproducibility study, the end result may be a low kappa value because of two predisposing factors: the overall agreement and the prevalence.
First, an overall agreement of less than 0.80 has the risk of resulting in a low kappa value.
Therefore, in the overall agreement period of the study (see figure 10, page 27), it is essential that observers try to achieve a substantial overall agreement P_o preferably above the level of 0.80. In this way the effect of the P_o on the final kappa value is under control.

Secondly, as shown above, very high and very low prevalences of the index condition result in low kappa values. Therefore we developed a theoretical method to influence the prevalence of the index condition in advance.

In figure 6 the prevalence/kappa curves are presented for the overall agreements P_o ranging from 0.83 till 0.98.

P_O	RANGE PREVALENCE
0.83	0.32 – 0.69
0.86	0.22 – 0.78
0.89	0.17 – 0.84
0.92	0.11 – 0.89
0.95	0.07 – 0.94
0.98	0.02 – 0.96

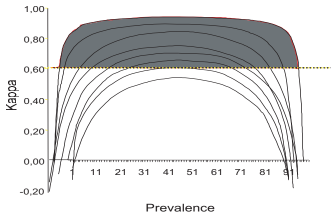


Figure 6. Kappa/prevalence curves of different overall agreements (0.83 – 0.98). The line through a kappa value of 0.60 demarcates the acceptable kappa area above this cut off line (gray area).

Note that the two lowest curves (P_o 0.83 and 0.86) are located beneath the line of the kappa value of 0.6. The curves with a P_o >0.90 have a substantial area (blue) above the 0.6 kappa cut off line.

To prevent unexpected low kappa values, because of unknown and too high or too low prevalences, we prefer to have a prevalence of the index condition near 0.50. The kappa values of prevalence of 0.50 are always located at the top of the curves.

Suppose that in the overall agreement period (see figure 10, page 27) we have achieved an overall agreement P_o of 0.85. We have 40 patients in whom we can study the reproducibility of a test.

Both Observer A as well as Observer B have each selected 20 patients, and each sends his/her 20 patients to the other observer. Each observer sends 10 patients who he judged to have a positive test and 10 subjects which he judged to have a negative test to the other observer. Based on an overall agreement of 0.85, both observers will agree in 85 % of the positive and negative judged tests. And disagree in 15 %. In figure 7 the scheme is presented.

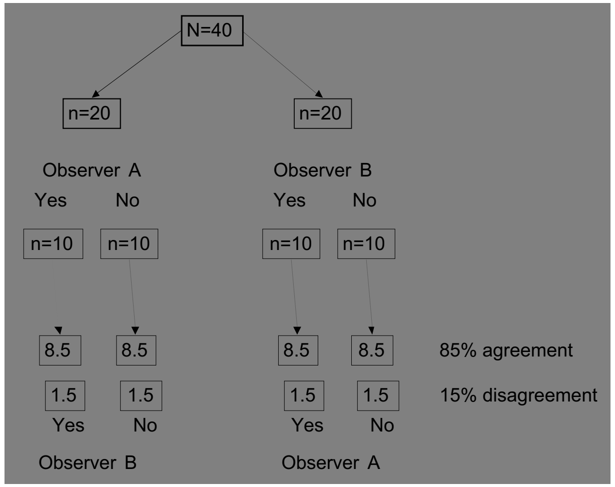
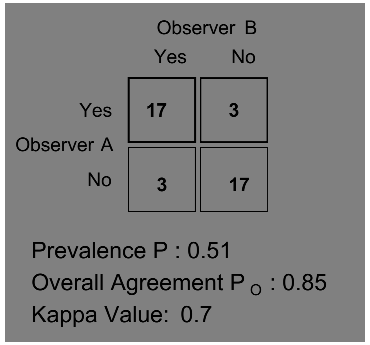


Figure 7. Scheme presenting the number of 40 patients with an overall agreement of 0.85, trying to get a prevalence of the index condition (positive test) of 0.50

Based on the number of patients in which the observers agree and disagree (figure 7), a kappa value can be calculated. In figure 8 a 2 x 2 contingency table shows the results. The prevalence is 0.50 with a overall agreement of 0.85, resulting in a kappa value of 0.70.



		Observer B	
		Yes	No
Observer A	Yes	**17**	**3**
	No	**3**	**17**

Prevalence P : 0.51

Overall Agreement P_O : 0.85

Kappa Value: 0.7

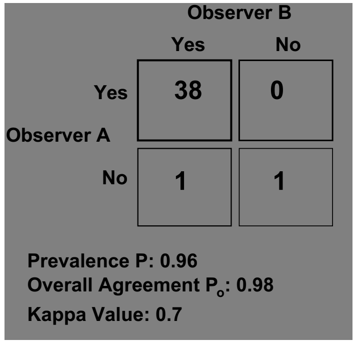
Figure 8. 2 x 2 contingency table based on the results of figure 7

By performing an overall agreement period in a reproducibility study with an overall agreement above the level of 0.80 and subsequently performing a procedure as illustrated in figure 7, one can influence the prevalence in advance resulting in a substantial kappa value of a test procedure. In a recent study, this proposed theoretical format was tested in practice and proved to be right (Patijn 2003, in press).

The easiest way of calculating the kappa value is to use a spreadsheet in which the formulae are integrated. In this way only the basic data has to be filled in and the kappa value is automatically calculated (see appendix 1). On the FIMM website a spreadsheet file can be downloaded.

III.10 Presentation Kappa Studies

In publishing the results of a reproducibility study, all aspects discussed under item 1 to 8 have to be presented. Furthermore, 2 x 2 contingency tables, the overall agreements and the prevalences are essential in a publication. In this way the reader of a paper can easily judge on what data the conclusion is based.
Figure 9 shows an example of a 2 x 2 contingency table. The calculation of the kappa value is also shown.



		Observer B	
		Yes	No
Observer A	Yes	38	0
	No	1	1

Prevalence P: 0.96
Overall Agreement P_o: 0.98
Kappa Value: 0.7

Figure 9. 2 x 2 contingency table of a reproducibility study of 40 subjects

III.11 References Kappa Literature

Barlo W, Lay M I, Azen P, A comparison of methods for calculating a stratified kappa, Statistics in Medicine, 1991; 10(9): 1465–1472

Cohen J, A coefficient of agreement for nominal scales. Educ Psychol Measurement, 1960;20:37–46

Cohen J, Weighted Kappa: Nominal Scale agreement with provision for scaled disagreement or partial credit, Psychol Bulletin, 1968: 70(4): 213–220

Cook R, Kappa, In: Encyclopedia of Biostatistics Eds. Armitage P, Colton T, Publishers Wiley, NY, 1998: 2160–2165

Deursen L L J M, Patijn J, Ockhuysen A L, Vortman B J, The value of different clinical tests of the sacroiliac joint, Ninth International Congress in Manual Medicine, London 1989, Abstr 16

Deursen L L J M, Patijn J, Ockhuysen A L, Vortman B J, The value of different clinical tests of the sacroiliac joint, J Manual Medicine 1990 (5): 96–99

Deursen van L L J M, Patijn J, Ockhuysen A L, Vortman B J, Die Wertigkeit einiger klinischer Funktionstests des Iliosakralgelenks, Manuelle Medizin 1992 vol 30(6): 43–46

Gjorup T, The Kappa Coefficient and the Prevalence of a Diagnosis, Meth Inform Med, 1988;27(4):184–186

Landis R J, Koch G G, The measurement of observer agreement for categorical data, Biometrics, 1977; 33: 159–174

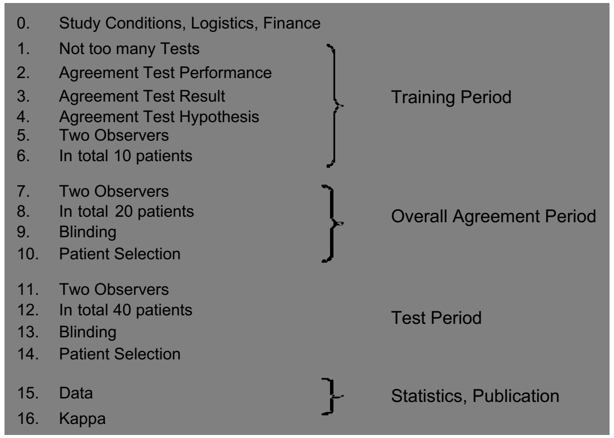
Lantz C, Nebenzahl E, Behaviour and Interpretaion of K Statistics: Resolution of the two paradoxes, J Clin Epidemiology, 1996; 49(4): 431–434

Patijn J, Stevens A, Deursen L L J M, Van Roy J, Neurological Notions on the Sacroiliac Joint, In Progress in Vertebral Column Research, First International Symposium on the Sacroiliac Joint: Its Role in Posture and Locomotion Editors A Vleeming, C J Snijders, R Stoeckart, Maastricht, 1991: 128–138

Patijn J, Brouwer R, Lennep van L, Deursen L, The diagnostic value of sacroiliac test in patients with non-specific low back pain, J Orth Med, 2000; 22(1): 10–15

IV. SEVEN GOLDEN RULES FOR A REPRODUCIBILITY STUDY

In figure 10 a scheme is presented of the different aspects and stages of a reproducibility study on which the Golden rules are based.
Reproducibility studies are easy to perform and not restricted to large institutes like universities. Private practices or other institutes with two or more practitioners in M/M Medicine are very suitable for this kind of study



0. Study Conditions, Logistics, Finance
1. Not too many Tests
2. Agreement Test Performance
3. Agreement Test Result
4. Agreement Test Hypothesis
5. Two Observers
6. In total 10 patients

Training Period

7. Two Observers
8. In total 20 patients
9. Blinding
10. Patient Selection

Overall Agreement Period

11. Two Observers
12. In total 40 patients
13. Blinding
14. Patient Selection

Test Period

15. Data
16. Kappa

Statistics, Publication

Figure 10. Plan of a reproducibility study

RULE 1 CREATE A CLEAR LOGISTIC AND RESPONSIBILITY STRUCTURE FOR THE REPRODUCIBILITY STUDY.
In a study one single person must be responsible for the entire process of the whole study.
This person is responsible for the logbook of the study. In this logbook all agreements and disagreements are written down and can be used as a reference cadre in group discussions. This person is responsible for the final updated format of the protocol. All participants have to sign this final protocol.

RULE 2 ALWAYS CREATE A TRAINING PERIOD BEFORE PERFORMING A REPRODUCIBILITY STUDY.
In the training period, it is essential for the future observers of a reproducibility study to discuss and define which tests and how many tests they are going to select for the reproducibility study. The decision on how many tests one wants to evaluate is dependent on the aim of the reproducibility study.

In the training period participants have to agree about the detailed performance of the test(s) that they are going to use for the reproducibility study.
20 patients can be used to discuss the precise sequence of procedure of the test(s). Finally, they have to agree about the precise performance of the test and make sure that each observer in a written protocol knows a standardised definition of the test procedure.

It is advisable not to restrict the agreed upon test procedure only to the patients of the study. But, by applying the same agreed upon test procedure to all one's patients visiting a clinic, it enhances the skills of the observers.
Participants have to agree how to define the outcome of the test(s) they are going to use for the reproducibility study. Participants have to perform the test(s) on the same 20 patients and to discuss the precise conclusions of the test(s). Finally, they have to agree about the precise judgement of the test and make sure that each observer in a written protocol knows a standardised definition of the test result. After every new decision, the logbook has to be updated.

Where a combination of tests is being studied, define the minimum number of positive tests for a final positive result of the test procedure.
Participants have to agree about the hypothesis of the test(s) they are going to use for the reproducibility study. Whatever test(s) selected for a reproducibility study, the observers have to investigate step by step the whole test procedure and agree about what the test really tests in their daily practice.

RULE 3 ALWAYS CREATE AN OVERALL AGREEMENT PERIOD BEFORE PERFORMING A REPRODUCIBILITY STUDY.
This period is essential to achieve a substantial overall agreement > 0.80. If the overall agreement is less than 0.80, participants have to discuss their agreements and have to pass the training period again.

RULE 4 ALWAYS USE A BLINDING PROCEDURE IN A REPRODUCIBILITY STUDY.
In the protocol it must be clear how the blinding is achieved not only with respect to the observers but also with respect to the patients. In most protocols, except with items such as pain, blinding is guaranteed when no information is exchanged either between observer and patient or between both observers. Use separate forms for each observer to record their findings.

RULE 5 ALWAYS DEFINE THE POPULATION FROM WHICH THE SUBJECTS ARE SELECTED.
This is essential to show how the selection was made (for example all patients on entrance) and no bias in selection of patients was performed.

RULE 6 ALWAYS MENTION THE DEFINITION OF THE SOURCE POPULATION, THE SELECTION METHOD, THE BLINDING PROCEDURE, THE DEFINITION OF TEST PROCEDURE AND TEST RESULTS IN MATERIALS AND METHODS WHEN PUBLISHING A REPRODUCIBILITY STUDY.

RULE 7 ALWAYS SHOW A 2 x 2 CONTINGENCY TABLE WITH THE PREVALENCE AND OVERALL AGREEMENT FIGURES IN RESULTS WHEN PUBLISHING A REPRODUCIBILITY STUDY.

V. VALIDITY

V.1 Gold or Criterion Standard

After achieving good reproducibility of a test procedure (the extent to which two observers agree about a test in the same population), the validity of a test has to be assessed.

Validity measures the extent to which the test actually does what it supposed to do. More precisely, the validity is determined by measuring how well a test performs against the gold or criterion standard. This is a major problem as well for diagnostics in gen-eral medicine as for diagnostics in M/M Medicine.
In M/M Medicine many characteristic diagnostic procedures, using for instance the end feeling in a passively performed test, are supposed to evaluate the mobility of the anatomical structure being examined. In the vast majority, only a hypothesis is available. For many tests in M/M Medicine, the gold or criterion standard has yet to be developed.

Two kinds of gold standards can be distinguished. First of all there is a gold standard for test procedures. For instances if a test procedure tests the range of motion, or resistance at the end of a passive motion, a gold standard has to be developed that measures in a quantitative way (degrees or N/cm^2) the range of motion or pressure in normal subjects. The evaluation of the quantitative method has also to include a test/retest procedure, to see whether the procedure shows the same data in the same normal subject on two different occasions.
In second instance, both the clinical test procedure and the quantitative method can be compared.

A second kind of gold standard for tests is related to the hypothesis of this test as taught by our teachers (SI-hypomobility) or with a diagnosis. This is the very problem as well for diagnostics in general medicine as for diagnostics in M/M Medicine.
The gold standard for a clinical test can be a radiological, a surgical finding, a post mortem, or a criterion based on data out of a normal population. So far, imaging techniques such as X-ray, CT and MRI are inconclusive in M/M Medicine, because a large number of normal subjects show abnormalities with these techniques.
In special cases, such as the Slump Test, which evaluates dural sac irritation for example from postoperative lumbar adhesions, MRI with gadolinium contrast can act as gold standard.
For some pain-provoking tests in M/M Medicine, the criterion standard is the effect of local anaesthesia in that particular area. The problem with this kind of criterion standard is that one is never sure about the systemic effect of local anaesthetics, and if we are dealing with a referred pain area, if were are sure that the pain is related to the anatomical structure we want to investigate, etc.

In M/M Medicine many tests are used to estimate the mobility of a joint by means of the end feeling. In this case two different policies can be followed. First, one can develop a quantitative method to evaluate the end feeling. In this case the end feeling procedure is validated clinically. Secondly, one can develop a quantitative method to estimate mobility of a joint. In this case, the mobility aspect of a clinical test is evaluated and therefore the real hypothesis of the test.
The list of above-mentioned examples is far from complete, but illustrates the way a gold standard can be developed.
In the absence of a well-defined criterion standard, sometimes a consensus view of experts using some other tests is used as a criterion standard. The problem with the consensus view is that the experts are only agreeing about a test procedure based on hypothesis and the real validity of a test remains uncertain.

In M/M Medicine, before spending much energy to defining gold standards, it is essential that first of all the test procedures are reproducible.

V.2 Sensitivity and Specificity

It has no sense in reproducibility studies to estimate the sensitivity and specificity, when no gold standard is available.

In sensitivity and specificity studies, 100 subjects are sufficient.
The same group of 100 patients is assessed with the test in question and with the gold standard (see 2 x 2 contingency table below). Cases **a** and **d** are correct, cases **c** and **b** are respectively false positive and false negative. A good test has to have few false-positive and false-negative results.
The prevalence of the index condition is illustrated by the formula: (a+c)/n.
It is essential to realise that the prevalence of an index condition can vary in different institutes, countries and from time to time.

The sensitivity of a test is defined as: the proportion of the cases that have the index condition **(a+c)** that the test correctly detects. In formula: **a/(a+c).**

The specificity of a test is defined as: the proportion of the cases that do not have the index condition **(b+d)** that the test correctly detects. In formula: **d/(b+d).**

Both sensitivity and specificity are needed to determine the validity of a test and always have to be presented together in a paper.

	Criterion Standard		
Result of Test	**positive**	**negative**	
positive	**a**	**b**	**a+b**
negative	**c**	**d**	**c+d**
	a+c	**b+d**	**n = a+b+c+d**

V.3 Positive and Negative Predictive Value

To translate the statistics of sensitivity and specificity figures to daily practice, the physician has to know in the individual patient the chances whether a positive test is truly positive as opposed to false-positive. This is expressed in the so-called "positive predictive value of a test". In the 2 x 2 contingency table above, the formula of positive predictive value of a test is: **a/(a+b)**. One has to realise that the positive predictive value of a test is dependent of the prevalence of the index condition **(a+c)/n.**

Suppose we have 1000 subjects with a sensitivity and specificity of respectively 0.8 and 0.7 and a prevalence of the index condition is 10% (see 2 x 2 contingency table above).

This means that when **n**=1000, then **a+c** = 0.10 x 1000 = 100.

In case of a given sensitivity **(a / (a+c))** of 0.8:

a / (a+c) = 0.8
}→ a/100 = 0.8 → **a** = 80
(a+c) = 100 }→ (80+**c**) = 100 → **c** = 20
(a+c) = 100

If **a+c** = 0.10 x 1000 = 100, n – **a+c** = **b+d** = 1000 – 100 = 900

In case of a given specificity **(d / (b+d))** of 0.7:

d / (b+d) = 0.7
}→ **d**/900 = 0.7 → **d** = 630
(b+d) = 900 }→ (630+**b**) = 900 → **b** = 270
(b+d) = 900

The positive predictive value of a test in this case is
a / (a+b) = 80 / (80 + 270) = 0.22

The negative predictive value of a test is likewise calculated:
c / (c+d) = 270 / (80 + 630) = 0.30

Where there is a larger prevalence of the index condition **(a+c)/n**, the positive predictive value of a test **a/(a+b)** also rises with the same sensitivity and specificity figures. Therefore, the positive predictive value of a test only reflects the prevalence of the index condition and not the property of the test itself.

V.4 Likelihood Ratio

For estimation of the predictive power of a test, independently of the prevalence of the index condition, the likelihood ratio has to be calculated. By definition the likelihood ratio in formula is:

$$\text{Likelihood ratio} = \frac{\text{Sensitivity}}{1 - \text{specificity}}$$

Tests with likelihood ratios close to 1 or <1 are completely useless for daily practice.

First, some remarks about this likelihood ratio and its use in calculating the diagnostic confidence odds.
Normally, we are accustomed to think of percentages like prevalence or true positive figures. The likelihood ratio does not operate on percentages, but on odds based on prevalence and diagnostic certainty.
Odds are the ratio of changes in favour of a condition versus the chances against that condition being present.
For example if a condition has a prevalence of 60 %, the prevalence odds of the test being correct is 60 : 40 = 3 : 2. These odds can be changed again into decimal terms. If the prevalence odds are 3 : 2, the chances in favour are 3/(3+2) = 0.6.

By mathematical calculation, the diagnostic confidence odds are calculated by multiplying the likelihood ratio and the prevalence odds.

[Prevalence odds] x [Likelihood ratio] = [Diagnostic confidence odds]

To illustrate the importance of a large likelihood ratio in relation to the prevalence of a condition, an example is shown.

Suppose a condition, has a prevalence of 60 % in your practice. Based on reproducibility and validity studies you know that the sensitivity is 0.8 and the specificity is 0.98.

Based on the formula: $\text{Likelihood ratio} = \frac{\text{Sensitivity}}{1 - \text{specificity}}$
the likelihood ratio is 40.
If a patient with a particular condition enters your practice, with a known prevalence figure of 40 %, the chance of having this condition is 60 %.

The prevalence odds in favour of having the condition are 6 : 4.
The odds for diagnostic confidence is 6/4 x 40 = 60.
Diagnostic confidence odds = 60 : 1.
Diagnostic Confidence is 60/60+1= 0.98 = 98 %.

This means that you have improved your confidence from 60 % to 98 %. This is a good test.

When calculating for the same prevalence of 60 %, but with a likelihood ratio of 0.6, the diagnostic confidence will be only 0.47 or 47 %. This is less than the chance of 60 % of having the condition for a patient when entering your practice. This is a bad test.

Published results of validity studies, trying to advise the daily practitioner which test he has to perform, and only mentioning sensitivity and specificity figures, are worthless. If one knows the prevalence of a certain condition, one can calculate, based on the likelihood figures, the diagnostic confidence.

APPENDIX 1

RELIABILITY of DIAGNOSTICS in M/M MEDICINE

		Observer B		Total
		Yes	No	
Observer A	Yes	a	b	a + b
	No	c	d	c + d
	Total	a + c	b + d	n

Number of subjects = **n**

Overall agreement $\mathbf{p_o} = \dfrac{\mathbf{a + d}}{\mathbf{n}}$

Expected chance agreement $\mathbf{p_c} = \dfrac{\mathbf{a + b}}{\mathbf{n}} \times \dfrac{\mathbf{a + c}}{\mathbf{n}} + \dfrac{\mathbf{c + d}}{\mathbf{n}} \times \dfrac{\mathbf{b + d}}{\mathbf{n}}$

Kappa = $\dfrac{\mathbf{p_o - p_c}}{\mathbf{1 - p_c}}$

Prevalence P = (a + [b + c]/ 2) / n

In a spreadsheet the following columns can be defined; see figure above:

Only data **a, b, c, d** has to be filled in:
Column A: data **a** (see 2 x 2 contingency table)
Column B: data **b** (see 2 x 2 contingency table)
Column C: data **c** (see 2 x 2 contingency table)
Column D: data **d** (see 2 x 2 contingency table)
Column E: data **n** Formula =A1+B1+C1+D1
Column F: data **a+b** Formula =A1+B1
Column G: data **a+c** Formula =A1+C1
Column H: data **c+d** Formula =C1+D1
Column I: data **b+d** Formula =B1+D1
Column E: data **a+d** Formula =A1+D1
Column K: Prevalence Formula =A1/E1+B1/2 x E1+C1/2 x E1
Column L: Overall Agreement P_o Formula =J1/E1
Column M: **(a+b)/n** Formula =F1/E1
Column N: **(a+c)/n** Formula =G1/E1
Column O: **(c+d)/n** Formula =H1/E1
Column P: **(b+d)/n** Formula =I1/E1
Column Q: data column M x N Formula =M1 x N1
Column R: data column O x P Formula =O1 x P1
Column S: Expected change agreement P_c Formula =Q1+R1
Column T: $P_o - P_c$ Formula =L1–S1
Column U: $1 - P_c$ Formula =1–S1
Column V: Kappa value Formula =T1/U1